FOR OFFICIAL USE ONLY

51

Return to

RESEARCH AID

THE PRESENTATION OF STATISTICAL DATA

ORR Project 11.3
11 March 1955

CENTRAL INTELLIGENCE AGENCY

OFFICE OF RESEARCH AND REPORTS

FOR OFFICIAL USE ONLY

WARNING

This material contains information affecting the National Defense of the United States within the meaning of the espionage laws, Title 18, USC, Secs. 793 and 794, the transmission or revelation of which in any manner to an unauthorized person is prohibited by law

Approved For Release 2010/11/26 : CIA-RDP07-00617R000100020001-9

FOR OFFICIAL USE ONLY

RESEARCH AID

THE PRESENTATION OF STATISTICAL DATA

CIA/RR RA

(ORR Project 11.3)

NOTE ON CLASSIFICATION

This research aid is FOR OFFICIAL USE ONLY, but all pages except the cover pages and the title page are unclassified.

CENTRAL INTELLIGENCE AGENCY

Office of Research and Reports

FOR OFFICIAL USE ONLY

Approved For Release 2010/11/26 : CIA-RDP07-00617R000100020001-9

FOREWORD

The following comments on handling and presenting statistical data are prompted by errors most commonly observed in reports submitted for publication. Most readers will be familiar with the procedures outlined. If the present research aid reminds some analysts of fundamentals forgotten, or if it helps a few to supplement their research background, it will serve its purpose.

Approved For Release 2010/11/26 : CIA-RDP07-00617R000100020001-9

CONTENTS

Tables

Illustrations

Approved For Release 2010/11/26 : CIA-RDP07-00617R000100020001-9

THE PRESENTATION OF STATISTICAL DATA

I. Proofing.

The most common errors in handling quantitative data are mistakes in simple arithmetical computation, transposition of figures, or other faulty transcription of numbers. Theoretically, such errors never should occur. In actuality, they will creep into even the most carefully done research. However, constant awareness of the danger of such errors, methodical research procedure, and careful proofing will keep them to a minimum in the final product. A few rules for proofing, well known by all research workers, but too frequently forgotten in moments of haste or carelessness, follow:

1. All data should be proofed carefully following any transcription and again when the project is completed.

2. Final proofing should be against original sources (whenever possible) to avoid perpetuation of transcription errors that may have developed during the course of research.

3. Proofing should be conducted by some one other than the person who did the original work, since it is possible to make the same mistake repeatedly.

4. If one must proof one's own work, an attempt should be made to reverse processes, reading from copy to original, adding from bottom to top columns originally added from top to bottom (or subtracting from the totals figures that originally were added), multiplying where division has been performed, and so forth.

5. Consideration of the general order of magnitude of figures is an essential part of the proofing process. Such consideration of the logic of figures, as presented, will guard against significant errors arising from failure to read a slide rule correctly or misplacing a decimal point.

II. Significant Numbers.

1. Approximate Nature of Significant Numbers.

Spurious accuracy, a common mistake in the handling of statistical data, too frequently is not recognized as an error. Numbers used in abstract arithmetical work mean exactly what they say. The number 7 (or 7.00 ... 0) means exactly seven, no more and no less. Unfortunately, however, data used in most statistical computations are not so precise. Except in cases where an actual count is possible, data are, at best, measurements, the accuracy of which is limited both by the quality of the instrument and by the accuracy of the observer.* Inevitably, some estimation is involved.

The Rand McNally Reference and Road Atlas for 1950 states that the distance from Washington, D.C., to Philadelphia is 141 miles. Timetables give the railroad distance between these two cities as 135 miles. Although distances by automobile or railroad can be measured by instruments, the chance that either of the distances given is accurate to the last foot, or even to the nearest tenth of a mile, is very slight. The last digit in either case is an approximation. What the last digits signify is that the actual railroad distance lies somewhere between 134.5 and 135.5 miles, while the highway distance lies somewhere between 140.5 and 141.5 miles. If one is interested only generally in the distance between the two cities (overlooking the fact that the distances by automobile and railroad are two different measurements), it would be adequate to say, on the basis of either measurement, that the two cities are approximately 140 miles apart, since both distances round to 140. In this case, only two digits are significant, and the second digit (4) is an approximation. The figure signifies that the true distance between Washington, D.C., and Philadelphia lies somewhere between 135 miles and 145 miles.

Significant terminal digits are approximations indicating that the true figure lies within plus or minus one-half of one unit

* What follows concerning the approximate nature of measurements should not be confused with counting. It is possible to count accurately -- that is, the exact number of people present in a room can be determined accurately by simple count. Such a datum, being accurate, may be expressed to any degree of accuracy compatible with companion data.

of the last digit. For example:

58 = 57.5 to 58.5
58.0 = 57.95 to 58.05
58.00 = 57.995 to 58.005

(See IV, below, for a discussion of procedure followed in rounding odd or even terminal digits.)

2. Calculation with Significant Numbers.

Since data which purport to be actual measurements are only approximations, it is obvious that data which have been estimated must be treated as approximations. A few practices commonly accepted in the handling of such approximations follow:

a. Write only as many digits as are known to be correct (recognizing that the last digit will be an approximation), and add as many zeros as are necessary to locate the decimal point.

b. Treat as significant all digits except zeros which are included to indicate the location of the decimal point. (Both 2,000 and 0.0002 contain one significant figure.)

The significance of zero sometimes is difficult to determine. In general, zeros are significant unless they occur

(1) At the extreme left of a number (in the number 0.02 the zeros are not significant), or

(2) At the extreme right of a number and to the left of the decimal point.

In the latter case, significance of the zeros must be determined from context or from the method of writing the number. When terminal zeros appear in a column, it is fairly safe to assume that numbers ending in zero are significant to the same place as other numbers in the column. In the following example, zeros to the right of the thousands column obviously are present only to locate the decimal point and are not significant. Zeros appearing in the thousands column for 1941 and 1944

would appear to be significant. It would be safe to assume that values for all years have four significant figures.

Year	$ Value
1941	5,500,000
1942	6,725,000
1943	7,805,000
1944	7,950,000
1945	8,891,000

The significance of terminal zeros is less obvious where a number is not related so clearly to other numbers which do not end in zero. The number 100,000 standing by itself might be regarded as having only one significant figure, yet one, two, three, or more of the zeros could be significant. In such cases, the number of significant figures can be indicated in a footnote or by writing the number in what is known as standard, or scientific, notation. Standard notation should be used only when it can be assumed that the reader will be familiar with such notation. When this cannot be assumed, standard notation may be used to keep decimals and significance of figures straight in the calculation stage but should be omitted from finished reports.*

* Standard notation is based on the principle that every number can be expressed as a multiple of some power of 10. An illustration of this principle follows:

$$
\begin{aligned}
15{,}625 &= 1.5625 \times 10^{4} \\
1{,}562.5 &= 1.5625 \times 10^{3} \\
156.25 &= 1.5625 \times 10^{2} \\
15.625 &= 1.5625 \times 10^{1} \quad \text{or } 1.5625 \times 10 \\
1.5625 &= 1.5625 \times 10^{0} \quad \text{or } 1.5625 \times 1 \\
0.15625 &= 1.5625 \times 10^{-1} \\
0.015625 &= 1.5625 \times 10^{-2} \\
0.0015625 &= 1.5625 \times 10^{-3}
\end{aligned}
$$

Positive exponents indicate the number of places the decimal point must be moved to the right. Negative exponents indicate the number of places the decimal point must be moved to the left. Inclusion of zeros in the number to be multiplied by a power of 10 will indicate the number of significant zeros in the number so expressed. Thus, in the case of the number 100,000 mentioned above, if only the 1 is significant, the

c. When adding or subtracting approximate numbers, round the answer so that its last significant figure will fall in the same column as the last significant figure of the original number having its last significant figure farthest to the left.* To facilitate computation, figures may be rounded prior to addition or subtraction so that the answer will contain one (or two) more place(s) than ultimately will be retained. For example:

Dollars	Thousand Dollars
360,000	360
25,107,500	25,108
25,320,000	25,320
	50,788

---

standard notation would be 1×10^5. If the number contains significant figures to the thousands column, it should be written 1.00×10^5. If there are four significant figures, the 100,000 should be written as 1.000×10^5.

* It should be noted that sometimes data are collected in such fashion as to make determination of the significance of figures impossible. In other cases, significance may be known, but strict adherence to the rules for calculation with significant numbers may result in a loss of information. For example, the distance from point A to point B may be known to be approximately 100,000 miles. The distance from B to C may be approximately 25,000 miles and that from C to D approximately 15,000 miles. If rules for adding significant numbers are followed in adding the three together, the total distance from A to D would be reported as 100,000 miles, implying a range of from 50,000 to 150,000 miles. However, it may be known that 140,000 miles (the sum of the three approximate distances) is much closer to the true total distance than is 100,000. In such a case, use of ranges (see III, below) will prevent loss of information. By writing each of the measurements as range numbers -- for example, 50,000 to 150,000; 24,500 to 25,500; and 14,500 to 15,500 -- it is possible to calculate an upper and lower limit for the total distance -- that is, 89,000 to 191,000 miles. A best estimate of the measurement should be made and the range of error indicated in real or percentage terms. In this instance the best estimate is 140,000 ± 51,000, or, after rounding the result so that it carries only one more place than the least accurate figure, 140,000 ± 50,000. (Another means of expressing this would be 140,000 ± 36 percent.) Also, see the footnote on p. 9, below.

Since the first and last numbers are significant only to the ten-thousands column, the sum should be rounded to 50,790,000.

d. When multiplying or dividing approximate numbers, round the answer so that it contains no more significant figures than did the original number having the fewest significant figures. To facilitate calculation, round off the number having the largest number of significant figures so that it carries one (or two) more significant figure(s) than does the number having the smallest number of significant figures. Two examples follow:

(1) $236{,}553 \times 125 = 236{,}600 \times 125 = 29{,}575{,}000 = 29{,}600{,}000$

(2) $\frac{236{,}553 \times 125}{2.2} = \frac{237{,}000 \times 125}{2.2} = 13{,}465{,}909 = 13{,}000{,}000$

III. Ranges.

1. Use Where Significant Numbers Lack Precision.

Significant figures do not always permit precise expression of accuracy. They reflect accuracy of measurement, or error, to one-half of one unit of the last digit. For example, the highway distance (141 miles) between Washington, D.C., and Philadelphia was measured to the nearest mile. This information, written as 141 miles, implies that the true distance is between 140.5+ and 141.5- miles.

Where accuracy (or error) cannot be reflected adequately by significant numbers, and such accuracy is desired, a range must be used. Significant numbers always can be expressed as a range, but the converse is not always true. For example, it may be known that the means of measurement used in determining the distance between Washington, D.C., and Philadelphia are accurate only to within 2 miles in measuring distances of this magnitude. In this case, writing the result as 141 implies greater accuracy than is warranted. The distance might be written as a significant number in the following ways:

a. 141 or 140.5+ to 141.5-

b. 140 (14×10^{1}) or 135 to 145

Since the true distance lies anywhere from 139 to 143 miles (141 plus or minus 2), expression of the distance as a significant number is misleading. None of the implied ranges exactly coincides with the true range of possibilities. The true range cannot be expressed as a significant number. In such cases, the number should be expressed as a range (for example, 139 to 143, 141 plus or minus 2, or 141 plus or minus 1.4 percent). Whenever possible, ranges should be stated in terms of a best estimate (that is, the figure within the range which is felt to be the best single representation of the measurement) and an estimated range of error stated in real or percentage terms. These values, the best estimate and the error term, should be expressed to a degree of accuracy compatible with the significance of the original data.

Where it is not necessary to the understanding or use of the datum to have it expressed to the highest degree of accuracy, it is advisable to express it in rounded form. (This rounding should be at least sharp enough to include the true upper and lower limits. In the case of the distance between Washington, D.C., and Philadelphia this would mean 140 miles written as 14×10^1.)

2. Calculation with Range Numbers.

a. Method 1.

One method of performing calculations with range numbers follows:

(1) Addition: Add lower limit to lower limit, and upper limit to upper limit.

(2) Subtraction: Subtract the upper limit of the subtrahend from the lower limit of the minuend, and the lower limit of the subtrahend from the upper limit of the minuend.

(3) Multiplication: Multiply lower limit by lower limit, and upper limit by upper limit.

(4) Division: Divide the lower limit of the dividend by the upper limit of the divisor, and the upper limit of the dividend by the lower limit of the divisor.

The discussion of range numbers thus far has been limited to ranges written as a to b, for positive values only. The procedure is changed to the extent of inverting and sign changing when negative values appear in the ranges. Since negative values are not likely to occur, their treatment has been omitted. Whether the values are positive or negative, all that is required is trying the possibilities and choosing the minimum and maximum values obtained. The omitted procedure merely eliminates the necessity for trial and error where negative values occur.

b. Method 2.

(1) Another means of calculation with ranges treats the measurement and the error in the form $x \pm e$, where x is a positive approximate number, or measurement, and e is the error. Illustrations of this method follow*:

(a) Addition: Add approximate number (x) to approximate number, and error (e) to error:

$$\begin{array}{r} x_1 \pm e_1 \\ x_2 \pm e_2 \\ \hline (x_1 + x_2) \pm (e_1 + e_2) \end{array} \qquad \begin{array}{r} 10 \pm 2 \\ 5 \pm 1 \\ \hline 15 \pm 3 \end{array}$$

(b) Subtraction: Subtract the subtrahend, and add the errors:

$$\begin{array}{r} x_1 \pm e_1 \\ -x_2 \pm e_2 \\ \hline (x_1 - x_2) \pm (e_1 + e_2) \end{array} \qquad \begin{array}{r} 10 \pm 2 \\ -5 \pm 1 \\ \hline 5 \pm 3 \end{array}$$

* e_1 and e_2 in the illustrations are in real terms. When the errors are expressed as percentages, they must be translated into real terms before calculation using Method 2.

(c) Multiplication: Pick extreme values of the possible products:

$$\frac{x_1 \pm e_1}{x_2 \pm e_2} \qquad \frac{10 \pm 2}{5 \pm 1}$$

$$x_1 x_2 \pm (e_1 x_2 + e_2 x_1) + e_1 e_2 \qquad 50 \pm [(2 \times 5) + (1 \times 10)] + (2)(1) = 50 \pm 22 \text{ or } -18$$

Other results that may be obtained by multiplying $(x_1 \pm e_1)(x_2 \pm e_2)$ lie between the desired minimum and maximum values.

(d) Division:

$$\frac{x_1 \pm e_1}{x_2 \pm e_2} = \frac{x_1}{x_2} \pm \frac{x_2 e_1 + x_1 e_2}{(x_2)^2}$$

$$\frac{10 \pm 0.2}{5 \pm 0.1} = \frac{10}{5} \pm \frac{(5 \times 0.2) + (10 \times 0.1)}{25} = 2 \pm 0.08 \text{ or } 2 \pm 0.1$$

This form is approximate and should be used only when the relative value of e is small.

(2) The discussion above has indicated how significant numbers and range numbers reflect error in results. Whether the error is handled as part of the calculation or as a separate calculation is of little importance. What is important is that the error be reflected in the result.*

* If the numbers used in demonstrating calculations with significant numbers are written as ranges and the computations redone in the manner indicated for range numbers, it will be discovered that the upper and lower limits so determined are, in most cases, outside the implied limits of the result expressed in significant numbers. This difference usually is ignored and the result expressed as indicated in II, 2, above on significant numbers.

The cumulative effect of such errors can seriously affect an end result. Assume that it is estimated that at the end of 1945 the gold-exporting country of Ruritania had a gold inventory of 25,000 ± 5,000 kilograms; that annual production during the period 1946-50 was 10,000 ± 2,000 kilograms; and that annual consumption, including trade, used 9,000 ± 3,000 kilograms. The gold inventory at the end of 1950 may be determined as follows:

$$
\begin{array}{r}
25{,}000 \pm 5{,}000 \\
+10{,}000 \pm 2{,}000 \\
+10{,}000 \pm 2{,}000 \\
+10{,}000 \pm 2{,}000 \\
+10{,}000 \pm 2{,}000 \\
\underline{+10{,}000 \pm 2{,}000} \\
75{,}000 \pm 15{,}000
\end{array}
$$

$$
\begin{array}{r}
-9{,}000 \pm 3{,}000 \\
-9{,}000 \pm 3{,}000 \\
-9{,}000 \pm 3{,}000 \\
-9{,}000 \pm 3{,}000 \\
\underline{-9{,}000 \pm 3{,}000} \\
30{,}000 \pm 30{,}000
\end{array}
$$

The cumulative effect of the error is to increase the possible range of error from 20 percent in 1945 to 100 percent in 1950. Thus in absolute terms the gold inventory may be anywhere from zero to 60,000 kilograms.

The error need not be greatest in the total. Data for a total figure may be more accurate than for the components, and the component figures, given and estimated, may carry greater error terms.

It should be emphasized that the best estimate may not be the midpoint of the range under consideration. In such circumstances the error term will not be symmetrical and will be in the form $75 \begin{smallmatrix} +\ 22 \\ -\ 18 \end{smallmatrix}$.

3. Confidence Intervals.

Whenever possible, ranges of error should be determined mathematically. However, such ranges should not include all possible values (100 percent probability). Inclusion of values at the extremes of the range usually will increase the width of the range greatly. The range of error should be one of 95 percent probability -- that is, sufficiently wide to include the true value 19 times out of 20. Put another way, the odds should be 19 to 1 that the true value falls within the range. More inclusive ranges rarely are desirable when dealing with imprecise data.

When ranges of error must be determined subjectively, they still should be on a 95 percent probability basis. In spite of the difficulty of arriving at a 95 percent probability level subjectively, and in spite of the fact that it is, of necessity, approximated, a conscious attempt to limit the range of error in this fashion will eliminate portions of the range where values are less likely to fall. For example, a normal distribution theoretically will have values extending infinitely in either direction. A range of all values (100 percent probability) for the distribution would have to extend from minus infinity to plus infinity, whereas on a 95 percent probability basis the range of values may be, for example, 45 to 55. The latter range indicates with a high degree of probability that the measurement in question lies within its limits and should be acceptable for most practical purposes. The former range, which includes an infinite number of values outside the limits of 45 to 55, with only 1 chance in 20 of occurrence, is meaningless under most circumstances.

IV. Rounding.

1. For the sake of accuracy, numbers should be rounded to eliminate all digits which are not significant.

2. For the sake of the reader, even significant digits should be eliminated when they are unnecessary for precise comparison or clarity of presentation.

3. If the figures to be rounded off amount to less than one-half of one of the units retained, the last digit retained will remain unchanged (for example, 425,499 would round to 425,000).

4. If the figures to be rounded off amount to more than one-half of one of the units retained, the last digit retained will be increased by one (for example, 425,501 would round to 426,000).

5. If the figures to be rounded off amount to exactly one-half of one of the units retained, the last digit retained will remain unchanged if it is even and be increased by one if it is odd (for example, 424,500 would round to 424,000, while 425,500 would round to 426,000). Treating zero as an even digit, there are 5 odd and 5 even digits. In an infinitely large number of cases this procedure will increase the terminal digit half of the time and leave it unchanged half of the time. Hence its application will tend to reduce the chance of positive or negative bias in rounding.

6. In calculation, use unrounded figures carrying one or two more digits than the number of significant figures which can be carried legitimately in the final answer.

V. Totals.

In presentation of rounded data, correctly rounded totals frequently are not exactly equal to the apparent sum of components. If it is feared that this will bother readers, totals may be footnoted to explain that the discrepancy is due to rounding. Another technique sometimes used is to force the total by adjusting the figures. This technique introduces inaccuracies which would be avoided by simply footnoting discrepancies. If it is used, the component figures, not the total, should be adjusted, and care must be taken not to misrepresent facts significantly. In forcing totals, adjustments should be made so that percentage changes in individual figures will be kept to a minimum.

VI. Index Numbers.

Index numbers are a useful device when one wishes to focus attention on relative changes in some quantity or quantities without considering the absolute amounts of the quantity or of the changes. Since index numbers usually are in percentage terms, they also are useful for comparing changes originally measured in noncomparable units (for example, physical and value terms).

1. Simple Relatives.

One of the simplest and most frequently used indexes is the simple relative, which is nothing more than the expression of each datum in a series as a percentage of some other datum in the series which has been chosen as a base and equated to 100 percent. This may be illustrated as follows:

Year	US Midyear Population (Thousand)	Index (1937 = 100)
1937	128,961	100
1951	154,360	120
1952	156,981	122
1953	159,696	124

The most common type of index is one that measures relative changes occurring over time. Consequently, the base selected usually is a measurement for some time period. The base selected will depend on what the index is to illustrate.

Simple relatives are sometimes expressed as percentages of the preceding year. For example, 1946 may be expressed as 85 percent of 1940, 1947 as 105 percent of 1946, and 1948 as 101 percent of 1947. Such relatives are known as link index numbers. Link index numbers may be converted to a common base, or chain index, by setting the base year equal to 100 and obtaining successive values by multiplying the link index number for each year by the chain index number for the preceding year. A chain index can be constructed from the link index numbers given in the example above as follows:

1940 = 100; 1946 = 85 (or 100 x 85); 1947 = 89 (or 85 x 105); 1948 = 90 (or 89 x 101). (See the following tabulation.*)

Sometimes it is desirable to shift the base of such relatives. There can be a number of reasons for doing so -- for example, desire to compare indexes originally on different bases, desire to focus attention on comparison of data with that for some date of special interest, or desire to splice overlapping indexes together. Shifting relatives from one base to another is done by setting the index value of the new base period (for example, a year) equal to 100 and determining the other values on the new base by dividing each old index value by the old value of the new base and multiplying by 100.** (See the following tabulation.*)

* P. 14, below.

** Such shifting cannot be done indiscriminately. Some indexes require recomputation if the base is to be shifted.

Year	Link Index Number	Multiply by	Chain Index Number (1940 = 100)	Chain Index Number (1950 = 100)
1940	100		100	109
1946	85	100	85	92
1947	105	85	89	97
1948	101	89	90	98
1949	104	90	94	102
1950	98	94	92	100
1951	99	92	91	99
1952	100	91	91	99
1953	104	91	95	103

In calculating percentage increases and decreases, a frequent error is failure to subtract 100, resulting in a figure expressed in terms of the base rather than as an increase over or a decrease from the base. Such calculations can be performed properly by dividing the base figure into the other, multiplying the quotient by 100, and subtracting 100. The resulting difference indicates the increase or decrease with appropriate sign: for example, if the base = 400,

$$\left(\frac{800}{400} \times 100\right) - 100 = +100 \text{ percent (increase)}$$

$$\left(\frac{300}{400} \times 100\right) - 100 = -25 \text{ percent (decrease)}$$

2. Aggregate and Weighted Indexes.

Frequently, expression of single values as percentages of a base is inadequate, and indexes become more complex, entailing aggregation and/or weighting.

With only a limited number of series of data, and knowledge of their appropriate weights, it is possible to reflect the activity of an entire economy. For example, by utilizing production data for four ferroalloying metals, representing 45 percent of Ruritania's total ferroalloying metals production, it is possible to construct an index reflecting production of all ferroalloys. The ferroalloying metals

index could then be used, together with similar production indexes for other types of metals, as a component index in constructing an all-metals index, which, in turn, could be used to construct an economy-wide index.

a. Simple Aggregates.

Simple aggregation requires only that the summation of data for one period be expressed as a percentage of a similar summation for the chosen base period. Where data are all in the same units, they may be added directly. Where not in the same units, data can be expressed as relatives (each datum as a percentage of the same datum in the base period), be added, and then be expressed as a percentage of the base period. Illustrations are given in Tables 1 and 2.*

Table 1

Production of Selected Ferroalloying Metals in Ruritania
1946-52

Thousand Metric Tons

	1946	1947	1948	1949	1950	1951	1952
Manganese	10.25	10.50	10.75	10.25	10.75	10.90	11.00
Molybdenum	0.48	0.51	0.55	0.55	0.58	0.58	0.58
Chromite	0.90	0.90	0.85	0.80	0.85	0.88	0.90
Tungsten	0.12	0.15	0.15	0.15	0.15	0.15	0.20
Total	11.75	12.06	12.30	11.75	12.33	12.51	12.68
Index (1948 = 100)	96	98	100	96	100	102	103

* Table 2 follows on p. 16.

Table 2

Indexes of Production of Selected Ferroalloying Metals in Ruritania
1946-52

1948 = 100

	1946	1947	1948	1949	1950	1951	1952
Manganese	95.3	97.7	100.0	95.3	100.0	101.4	102.3
Molybdenum	87.3	92.7	100.0	100.0	105.5	105.5	105.5
Chromite	105.9	105.9	100.0	94.1	100.0	103.5	105.9
Tungsten	80.0	100.0	100.0	100.0	100.0	100.0	133.3
Total	368.5	396.3	400.0	389.4	405.5	410.4	447.0
Index	92	99	100	97	101	103	112

Table 1 illustrates the aggregation of actual values (where the data are all in the same units) and demonstrates the domination of the index by the highest weight. Table 2, where the data are expressed in relatives, eliminates this domination by a single item.* Use of relatives has the added advantage of reducing all data to unitless terms, hence making possible the addition of data that originally are not in comparable, or like, units. In both instances, however, the relative importance of the four metals (chosen as being representative of the entire population of ferroalloying metals) has not been given any consideration. In Table 1, manganese dominates the index merely because it has the greatest weight of production and not because of any measure of the relative value of its production to the economy. In Table 2 the four components have been assigned equal weights -- that is, all are treated as being of equal importance.**

* It should be noted that an arithmetic mean was used in arriving at the annual index in this case. Other measures of central tendency could have been used. Each has its advantages and disadvantages.

** Indexes such as those in Tables 1 and 2, where, without recourse to additional information, all components are treated as being of equal importance, usually are called unweighted, or simple aggregate, indexes.

b. Weighted Aggregates.

Frequently data must be weighted to avoid unreasonable domination of the index by certain of the components and to insure that all components exert an influence proportionate to their relative importance to the economy. Suppose that ferroalloying metals were priced (in Ruritanian macropounds) as follows:

Metal	Price per Metric Ton (1947-49 Average)
Manganese	0.05
Molybdenum	10.30
Chromite	0.30
Tungsten	12.10

If these prices are used as weights in constructing an index from the data contained in Table 1, the resulting index will better reflect each metal's share (of the value) of production. In Table 3 the production data from Table 1 have been multiplied by the price data given above.

Table 3

Production of Selected Ferroalloying Metals in Ruritania
(Weighted by 1947-49 Average Prices)
1946-52

Thousand Macropounds

	1946	1947	1948	1949	1950	1951	1952
Manganese	0.51	0.52	0.54	0.51	0.54	0.54	0.55
Molybdenum	4.94	5.25	5.66	5.66	5.97	5.97	5.97
Chromite	0.27	0.27	0.26	0.24	0.26	0.26	0.27
Tungsten	1.45	1.82	1.82	1.82	1.82	1.82	2.42
Total	7.17	7.86	8.28	8.23	8.59	8.59	9.21
Index (1948 = 100)	87	95	100	99	104	104	111

The production index from Table 3 indicates a sharper rate of growth than do the indexes from Tables 1 and 2. The weights have given the data a new perspective, and the dominating component of the index is now molybdenum. The weighted index numbers reflect the production of ferroalloying metals more accurately than do the indexes from Tables 1 and 2, since the weights introduce the economy's relative evaluation of the different metals.

As the weighted index is continued for several more years, it is possible that it will reflect production less realistically. Changes in demand, prices, or other factors may make the weights derived from 1947-49 data no longer suitable. This sort of bias becomes more likely as time extends farther and farther from the base, and it may become necessary to change both the base period and the system of weights.

There are other types of indexes that might be constructed. Different weights, weights that vary from year to year, and weighted relatives are but three of many possibilities. These other indexes also would have their biases and would differ to some extent in what they measure. Care must be used in selecting the index to be used so that it will best reflect what is to be indicated. The data used must be appropriate and must be combined in the most suitable manner. The base chosen should minimize bias. Consideration should be given to the tendency of bias to increase as the span of time moves farther from the base. The brevity of the present research aid precludes more than hinting at many problems which arise in working with index numbers. The reader is urged, therefore, to consult detailed literature on index numbers for further information.

VII. Computing Rates of Increase or Decrease.

Sometimes it is desirable to compute the average rate of increase or decrease during a specified time period. A common error committed in computing such average rates of increase is the simple averaging of the increases from period to period. This procedure does not take into account the compound nature of the problem. Average rates of increase should be computed as follows:

$\frac{A}{B} = (1 + r)^n$, where A = final amount, B = base or initial amount, r = rate of increase for each time period, and n = number of time periods of increase or decrease.* Suppose that the price index of a commodity is as follows:

Year	Index
1950	100
1951	100
1952	280
1953	300

Then A = 300, B = 100, n = 3. Hence $\frac{300}{100} = (1 + r)^3$. (It is apparent that the only values necessary are those for the base period, the final period, and the span of time.) This equation may be solved by dividing the logarithm of $\frac{A}{B}$ (3) by n (3) and subtracting 1 from the antilog: $\frac{\log 3}{n} = \frac{0.477121}{3} = 0.159040$; the antilog of 0.159040 = 1.442 = (1 + r); hence r = 0.44. The annual rate of increase from 1950 through 1953 is 44 percent per year. This indication of the trend is quite different from the result of 62 percent which would be obtained by averaging the annual increases as follows: $\frac{0 + 180 + 7}{3} = 62$.

VIII. Tabular Presentation.

A table should be a completely self-explanatory unit, whether it occupies part of a page, a whole page, or several pages. Tables introduced into the body of a report should, however, be introduced at a logical point, and should be referred to (by number) at that point, in the text. Accompanying textual commentary should highlight important portions of the table and emphasize conclusions based on it.

* This is the compound interest formula usually expressed as $\frac{\text{Final Amount}}{\text{Principal}} = (1 + r)^n$. Final Amount = Principal + Interest.

Commentary should not merely repeat data presented in the table. The following comments on tabular presentation should be considered in conjunction with Table 4.*

1. Numbering.

a. Tables should be numbered consecutively, in order of physical location. Where tables appear in appendixes, the first appendix table should bear the number following that of the last table in the text.

b. The table number should be above the title and centered on the page.

c. Tables should be referred to by number in textual commentary. (Where textual references are widely separated from the table, page numbers should be given in footnotes.)

d. Small tabulations within the text, which are not set up as tables, need not be numbered.

2. Title.

a. The title of a table should be in topic form, briefly indicating what, where, and when, in order of importance. For example, if a report deals with the harmonica industry in East Germany, a table emphasizing production might be headed "Production of Harmonicas in East Germany, 1930-52." If, on the other hand, the report is on musical instrument production in East Germany, a table on production of harmonicas might be headed "Harmonica Production in East Germany, 1930-52." If the place is to be emphasized, in a report on world production of harmonicas, the table might be headed "East German Production of Harmonicas, 1930-52." Whichever order, or emphasis, is chosen, it should be used consistently throughout a single report, or, in some instances, throughout a section of a report having a specific emphasis.

* Table 4 follows on p. 21. Note that this type of footnote should be inserted when the table does not follow on the same page as the main reference thereto (for an additional example, see p. 15, above). All other footnote references to tables should be by page number, with "above" or "below" added, as the case may be (for example, see p. 25, below).

Table 4

[TITLE] Production and Cost of Turtle Food in Selected Counties of Ruritania a/

[PREFATORY NOTE] (Excluding Cottage Production)

1938, 1946-53

[CAPTION]

[UNITS]

[STUB] Year	X b/ Production (Pounds)	X b/ Cost c/ (Dollars)	Y Production (Pounds)	Y Cost (Dollars)	Z Production (Pounds)	Z Cost (Dollars)	Total Production (Pounds)	Total Cost (Dollars)
1938	400	82	120 d/	22	Negligible	0.1 e/	520	100
1946	210	60	30	9.5	Negligible	0.1	240	70
1947	260	100	35	12	2.2	0.9	300 f/	110
1948	300	120	55	19	4.0	1.9	360	140
1949	300 g/	130	70	25	2.0	1.4	370	160
1950	320	150	90	35	Negligible	0.2	410	190
Total, 1946-50	1,400	560	280	100	8.2	4.5	1,700	660
1951 h/	350	160	100	43	3.3	3.0	450	210
1952	400	200 i/	120	52 j/	4.1	3.9	520	260
1953 k/	410	210	130	58	4.2	3.9	540	270
Total, 1946-53	2,600	1,100	630	250	20	15	3,200	1,400
Average, 1946-53	320	140	79	32	2.5	1.9	400	170

[BODY]

a. Except where indicated otherwise, all data contained in Table 4 are from source 103/. All data are rounded to two significant figures. Totals and averages are derived from unrounded figures and do not always agree with rounded data shown.
b. As of 1 July.
c. Except where indicated otherwise, cost data for X are from source 104/.
d. 105/
e.
f.
g.
h.
i.
j.
k. Data for 1953 are from the following sources: County X, 106/; County Y, 107/; County Z, 108/.

b. Dates showing the period covered should be placed on a separate line beneath the descriptive part of the title. They should be presented as follows: 1948, to indicate a single year; May 1948, to indicate a single month; 1936-37 (but 1899-1900), to indicate two consecutive years; 1936-40 (but 1895-1910), to indicate more than two consecutive years; Selected Years, 1940-50, to indicate scattered years within a period; 1895, 1900, and 1940-48, to indicate widely scattered years. Fiscal years, crop years, trade years, or averages of consecutive years should be presented as follows: 1945/46, not 1945-46. The type of year used, if other than a calendar year, should be indicated in the title of the table or in a footnote.

c. Titles should appear in initial capitals, should not be underscored, and should not be followed by a period.

d. When a table covers more than one page, the word "Continued" in parentheses should appear under the title on all pages except the first.

3. Prefatory Note.

A prefatory note may be placed directly beneath the descriptive part of the title and before the date(s) for the purpose of clarifying or limiting the title, provided this explanation can be given in a brief phrase. Such brief notes should be of a general nature, applying to all or most of the table. Longer explanations or explanations of specific items in the table always should be given in footnotes.

4. Spacing.

Double-space between the title (including the word "Continued" in parentheses, when used) and the beginning of the table. The unit of measurement, placed as indicated under Units (see 5, a, below), would mark the beginning of the table for this purpose.

5. Units.

a. Where the unit of measurement is the same throughout the table, it should be placed at the extreme right on the solid line marking the beginning of the table.

b. If there is more than one unit of measurement and the data are arranged in columns, units should be indicated in parentheses in the caption, below column headings. Where data are arranged in rows, a units column may be used to designate the unit of measurement for each row. Units in a units column are not to be enclosed in parentheses.

perhaps skip

6. Caption (Column Headings).

a. Column headings should be brief and in the singular.

b. Comparable column headings should be consistent -- for example, value is comparable to quantity, not tons.

c. Units of measurement should appear in parentheses under the appropriate heading (see 5, b, above).

d. Underscoring of column headings or subheadings should extend to the limits of all columns under the heading.

7. Stub (Row Headings).

a. Items in the stub (side entries) should be listed in the order best suited to the data -- for example, according to importance or geographically, alphabetically, and so forth.

b. If a second line is required for a stub entry, the second line should be indented, and related column entries should be placed opposite the bottom line of the stub entry.

c. Subheadings should be double-spaced below main headings and indented two spaces. Items subordinate to subheadings should be double-spaced below the subheadings and indented two spaces.

d. Totals and averages should be double-spaced below the entry which they follow, and the designation should be indented two spaces.

e. Entries in the stub should not be underscored.

8. Body.

a. Ciphers (0) should be entered where data have values of zero.

b. Where data have values equal to or less than one-half of the minimum unit being carried, the word "Negligible" should be entered.

c. Where data are not available, but the phenomenon which the data would represent is known to exist, enter "N.A." in the appropriate column.

d. Where data are not available and it is not known whether any item or activity exists to be represented, enter "Unknown."

e. Ditto marks, hyphens, and dashes should never be used in a table.

f. In figures of four or more digits the comma should be used.

g. Totals and averages should be underscored with a single line, grand totals and averages with a double line.

9. Footnotes.

a. Lower-case letters should be used as references to table footnotes.

b. In making footnote references, each line should be considered in its order, with more than one reference on a given line lettered consecutively from left to right.

c. Footnote references should be placed after an entry. When there is no entry, the footnote reference should be placed in the position of the entry.

d. When a table is more than one page long, footnote entries should appear at the end of the table. In such cases the first footnote reference should be followed by an asterisk (a/*), and the asterisked reference at the bottom of the first page should read, for example, as follows: "Footnotes to Table 3 follow on p. 6."

10. Sources.

Numerical source references should be kept out of the body of the table. References, numbered in the same numerical sequence used in the text, should follow title, caption (column headings), or row headings. Where use of multiple sources necessitates some clarifying

statement in addition to simple source references, lower-case letters may be substituted for reference numbers in the title, caption (column headings), or row headings, or, for individual items, in the body of the table with the reference number appearing in the footnote below (see Table 4*). The Sources appendix should carry source references for each entry.

11. Small Tabulations within the Text.

a. Small tabulations within the text need not have the format of a table and may be considered part of the text.**

b. Such tabulations always are introduced verbally, usually with a statement such as "Production of these items during the 2-year period was as follows:"

c. Footnotes to small tabulations within the text should be considered as text footnotes and treated accordingly.

IX. Graphic Presentation.

1. General Notes.

a. Graphic presentation is intended to reveal relationships of data at a glance. Graphs, or charts, should be completely self-explanatory units. Graphs often are numbered serially (as Figure 1, Figure 2, and so on), and whether numbered serially or not, they always should be introduced and referenced in the text.***

b. The title should indicate what, where, and when, in order of importance.

c. No source reference need be given on the graph itself. Source references should be given immediately below the graph or following any title or footnotes below the graph.

d. The vertical axis (ordinate) should indicate units of the dependent variable.

* P. 21, above.
** See pp. 3, 4, 5, 13, and so on, above.
*** See pp. 26, 27, and 28, below. Note that graphs will not be given page numbers. When they follow the last page of the report, they may be referenced as "inside back cover."

e. The horizontal axis (abscissa) should indicate units of the independent variable. Frequently, time is represented on the horizontal axis. In such cases the value plotted against time (where the unit of time is a period -- for example, year or month -- rather than a specific point in time) should be plotted at the midpoint of the particular time interval.

f. The legend should indicate what each curve, bar, and so forth, on the graph represents.

g. Footnotes should appear below the graph, on the left. When the graph is based on data set forth elsewhere in tables or text, it is not necessary to repeat either source references or explanatory notes. Reference to the table, or to the page on which the data originally are explained, is adequate. For example: "Data are from Table 1, p. 16, above" (or, as the case may be, "Data are from p. 16, above").

h. The zero point always should be indicated on arithmetic scale graphs. Where space limitations make inclusion of the complete scale, starting with zero, inconvenient, a break is used to indicate omission of part of the scale.

2. Arithmetic Graphs (Figure 1*).

a. Arithmetic graphs -- that is, those plotted on arithmetic scales -- are best for examination of series in real terms. From Figure 1 it may be seen readily that the values of curve A are at least 10 times as large as those of curve B.

b. If relative changes are to be examined, the arithmetic graph may lead to erroneous conclusions. For example, the fluctuations in curve A appear more violent than those in curve B. However, the fluctuation of curve A from 1936 to 1937 is 100 percent ($\frac{30 \text{ to } 60}{30} \times 100$), and of curve B for the same period 400 percent ($\frac{1 \text{ to } 5}{1} \times 100$). Because the purpose of graphics is to make relationships of principal importance instantly apparent to the reader, other forms of graphs are more suitable for indicating relative change.

3. Semilogarithmic Graphs (Figure 2*).

a. The semilogarithmic graph is a common method of charting relative change. Since equal distances on a logarithmic scale represent

* Following p. 28.

equal ratios, or percentage changes (as contrasted with equal amounts on arithmetic scales), the reader can ascertain relative change easily. For example, in Figure 2 relative growth between 1936 and 1940 obviously is greater for curve B than for curve A, since the vertical distance between values is greater for curve B than for curve A. This was not apparent when the same data were plotted on an arithmetic scale in Figure 1.

b. In logarithmic graphs, regardless of the scales used, equal distances always will mean equal ratios. (This makes possible the use of two or more scales on a single chart to facilitate comparison on a relative basis.) Because of this property, the semi-logarithmic graph is useful in illustrating ratios, comparison of data in different units, and examination of relative growths and differences.

c. A logarithmic scale never starts at zero, because each phase or cycle of the scale has values 10 times those of the corresponding values of the preceding cycle. If the scale were started at zero, each succeeding cycle also would start at zero (10 x 0). Although the initial value of the first cycle can be any value other than zero, proper selection of an initial value prevents having to use a scale with awkward units.

4. Graphs of Index Numbers (Figure 3*).

a. Another means frequently used to illustrate relative change is the graphing of index numbers. Where index numbers are expressed as percentages of a base (original units being lost in computation of the index numbers), the graphing of series of indexes makes possible quick comparison of relative changes in terms of the bases. (Equal distance reflects equal percentage change.) From Figure 3 it can be seen that for any given time period, the curve farthest from the base line of 100 has experienced the greatest change relative to the base.

b. Graphs of index numbers also make possible easy comparison of series measured in different terms and series that are considerably different quantitatively.

* Following p. 28.

5. Bar Charts (Figure 4*).

a. Bar charts may be used to plot frequency distributions (as are frequency polygons, or line charts, of the type described above). Bar charts are best suited for comparing data for a few years only, line charts being preferable when the number of items being compared and/or the independent variable (for example, time period) assumes more than a few values.

b. Bars should be of equal width, length being the only variant.

6. Pie Charts (Figure 5*).

Pie charts frequently are used to give a quick, rough comparison of relative sizes. Approximate percentages of the sections should be indicated on the diagram.

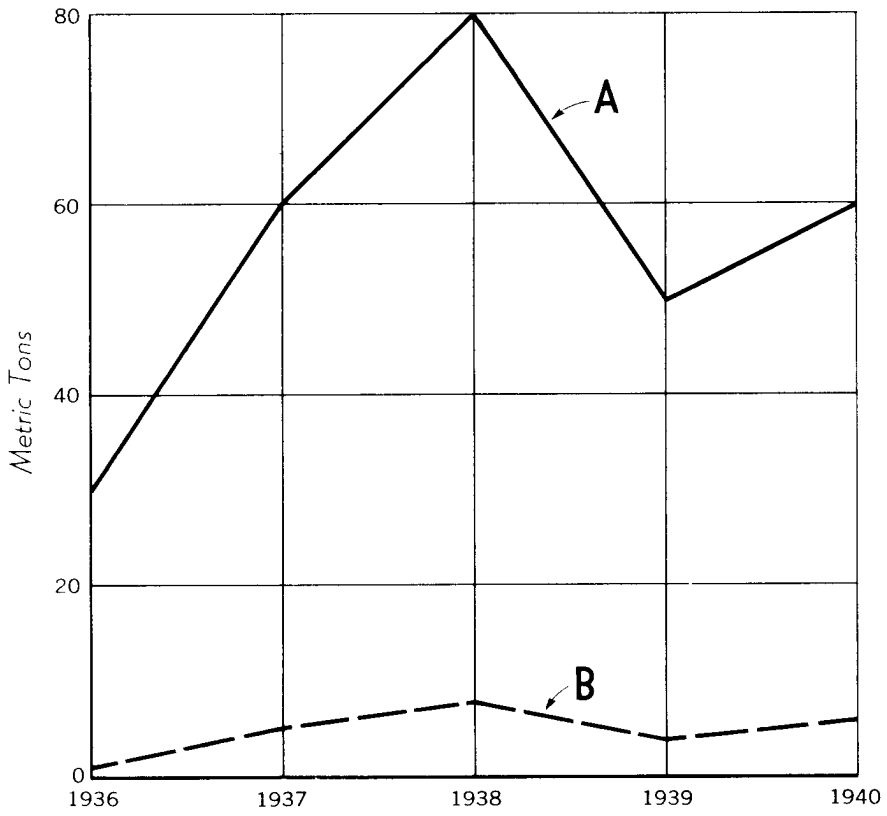
7. Pictorial Diagrams (Figure 6*).

Pictorial diagrams are a device for attracting attention. They are a rough means of presentation and may take almost any form. It usually is preferable to use items of constant size, varying the numbers, rather than to use items of different sizes to reflect different values. The latter approach leaves too much interpretation to the reader.

* Following p. 28.

Figure 1

RURITANIA
PRODUCTION OF A AND B*
1936-40

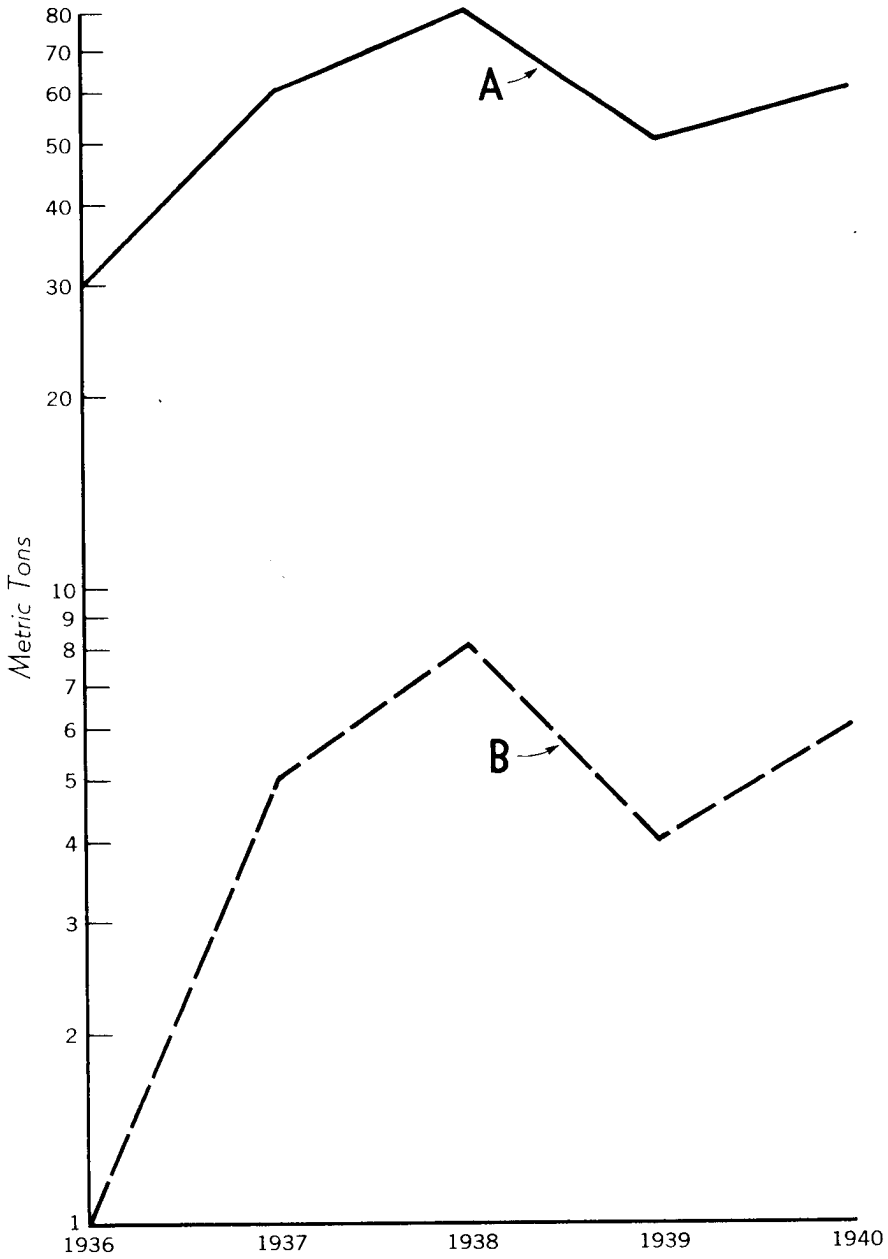


* *Excludes cottage production*

Approved For Release 2010/11/26 : CIA-RDP07-00617R000100020001-9

Figure 2

RURITANIA
PRODUCTION OF A AND B*
1936-40



* Excludes cottage production

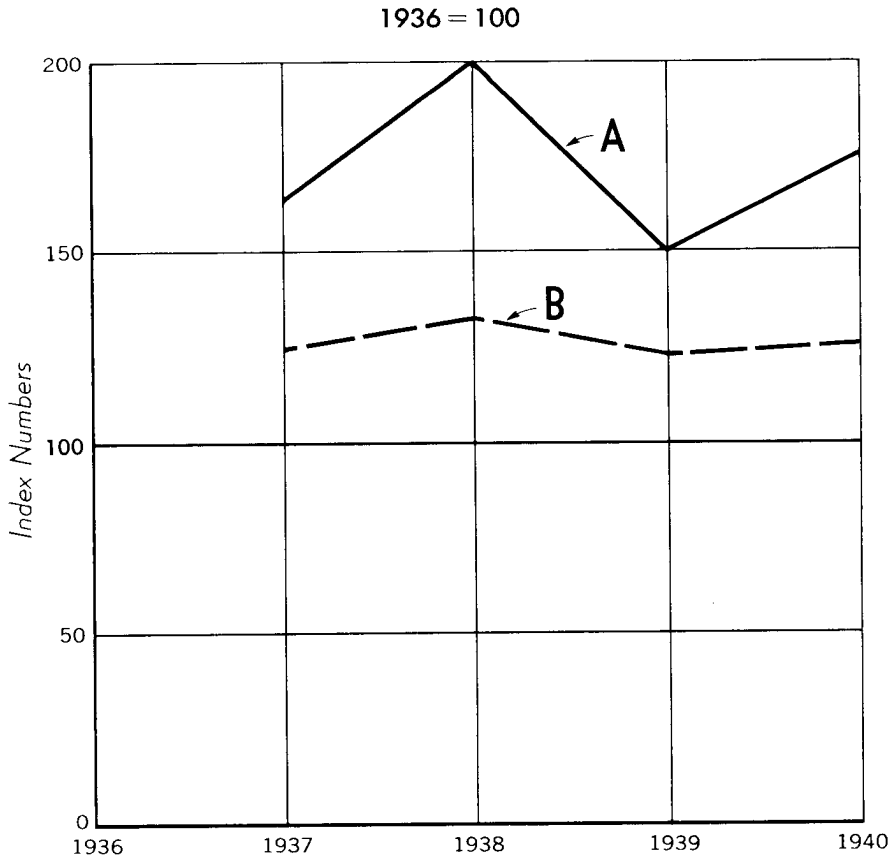
Approved For Release 2010/11/26 : CIA-RDP07-00617R000100020001-9

Figure 3

RURITANIA

INDEXES OF PRODUCTION OF A AND B*

1936-40



*Excludes cottage production

Approved For Release 2010/11/26 : CIA-RDP07-00617R000100020001-9

Figure 4

RURITANIA

PRODUCTION OF C AND D, 1935-37

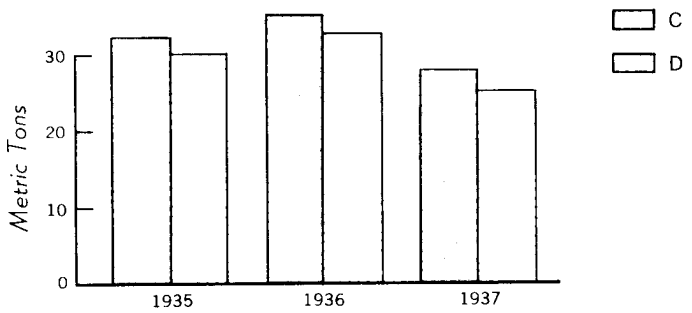


Figure 5

AREA OF RURITANIA, BY USE

1954

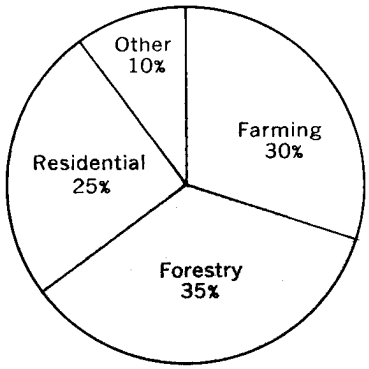
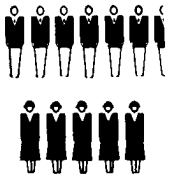


Figure 6

POPULATION OF RURITANIA, BY SEX

1954

(Each figure represents one million Ruritanians)



MALES

FEMALES

Approved For Release 2010/11/26 : CIA-RDP07-00617R000100020001-9

APPENDIX

BIBLIOGRAPHY

1. Arkin, Herbert, and Colton, Raymond R., Statistical Methods As Applied to Economics, Business, Psychology, Education, and Biology (College Outline Series), 4th edition, revised, New York, Barnes & Noble, Inc., 1950.
2. Croxton, F.E., and Cowden, D.J., Applied General Statistics, New York, Prentice-Hall, Inc., 1939.
3. Dwyer, P.S., Linear Computations, New York, John Wiley & Sons, Inc., 1951.
4. Fisher, Irving, The Making of Index Numbers, Boston, Houghton Mifflin Co., 1927.
5. Huff, Darrell, How to Lie with Statistics, New York, W.W. Norton & Co., Inc., 1954.
6. Key, V.O., Jr., A Primer of Statistics for Political Scientists, New York, Thomas Y. Crowell Co., 1954.
7. Mitchell, W.C., Index Numbers of Wholesale Prices in the United States and Foreign Countries, US Bureau of Labor Statistics, Bulletin No. 284, Washington, D.C., US Government Printing Office, 1921.
8. Tippett, L.H.C., Statistics, London, Oxford University Press, 1943.
9. Walker, Helen M., Mathematics Essential for Elementary Statistics, revised edition, New York, Henry Holt & Co., 1951.
10. Waugh, A.E., Elements of Statistical Method, 3d edition, New York, McGraw-Hill Book Co., Inc., 1952.